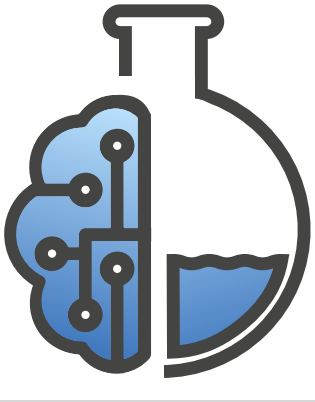


# IBM RXN for Chemistry

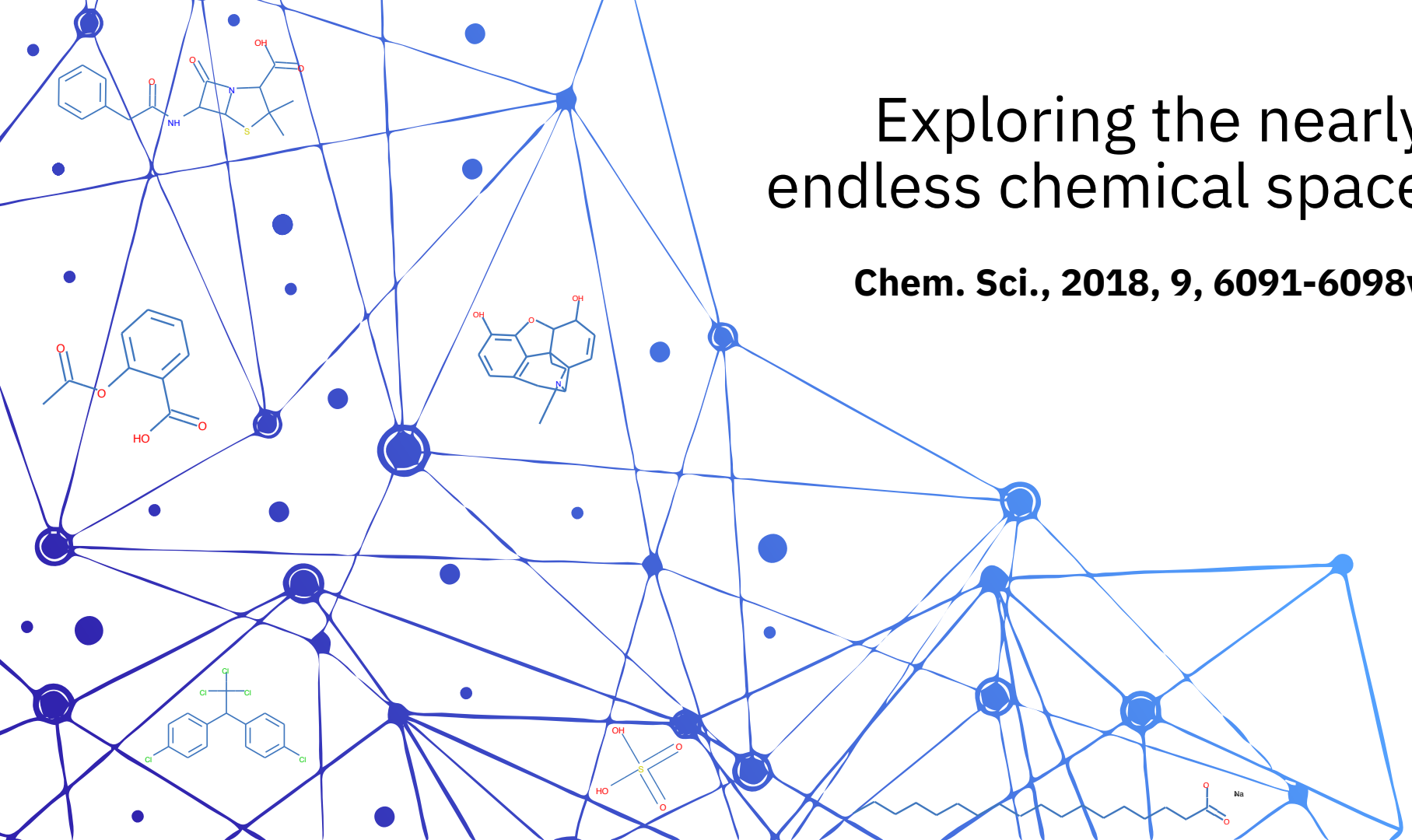
Theophile Gaudin, Philippe Schwaller, Riccardo Pisoni, Riccardo Petraglia, David Lanyi, Costas Bekas and

Teodoro Laino

IBM Research - Zurich, Switzerland

teo@zurich.ibm.com

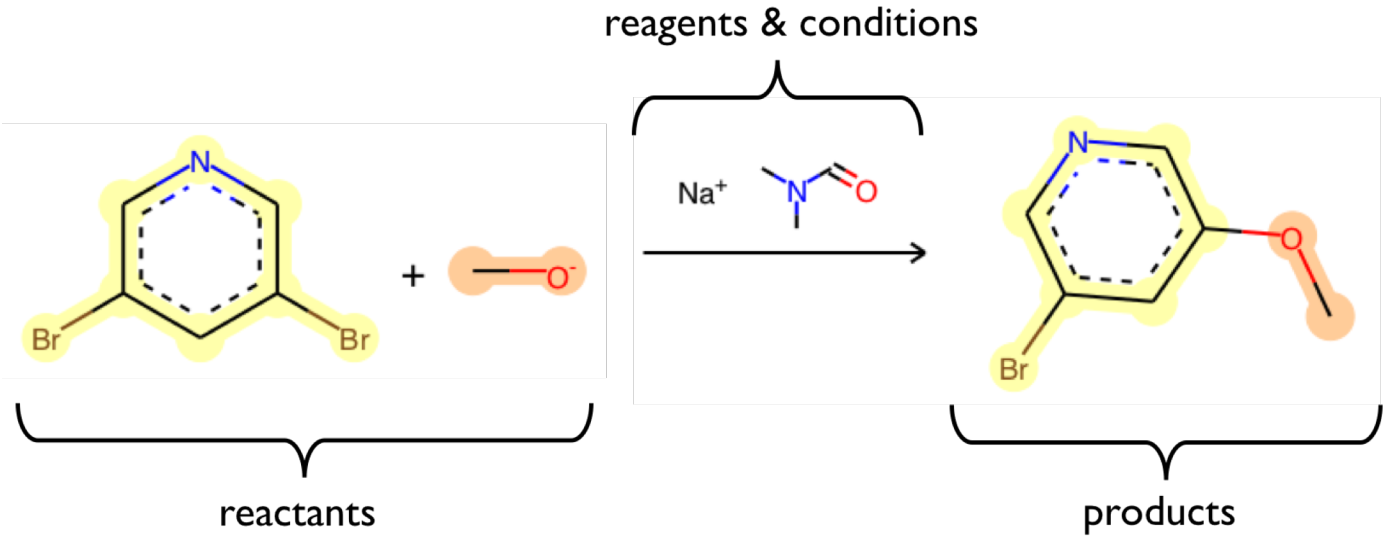
@teodorolaino



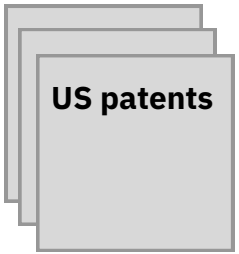
# Exploring the nearly endless chemical space

**Chem. Sci., 2018, 9, 6091-6098v**

# Chemical reaction prediction



# Data




Lowe (2012, 2017)

text-mining

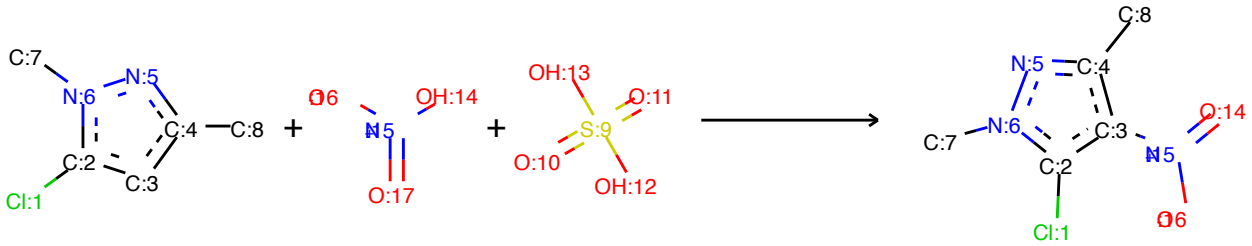
SMARTS
- reactants>>products
- partly correct atom-mapping
- >1M reactions

Jin et. al. (NIPS 2017)

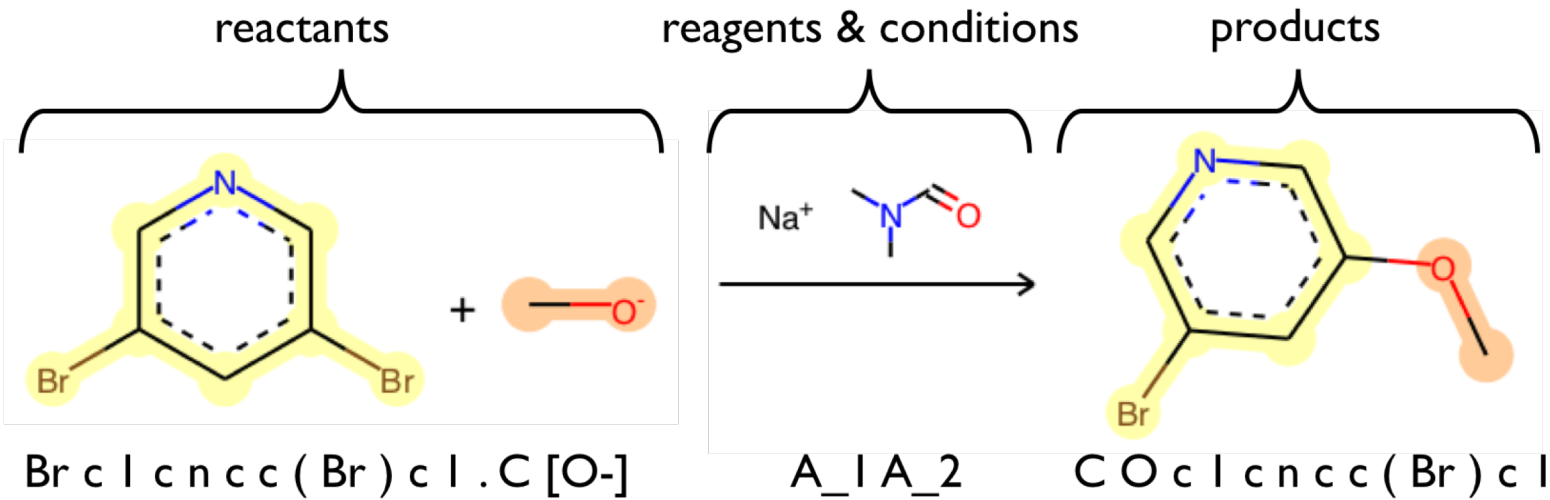
filtering

**Benchmark dataset (USPTO_500k)**

[Cl:1][c:2]1[cH:3][c:4]([CH3:8])[n:5][n:6]1[CH3:7].[OH:14][N+:15]([O-:16])=[O:17].[S:9](=[O:10])(=[O:11])([OH:12])[OH:13]>>[Cl:1][c:2]1[c:3]([N+:15](=[O:14])[O-:16])[c:4]([CH3:8])[n:5][n:6]1[CH3:7]



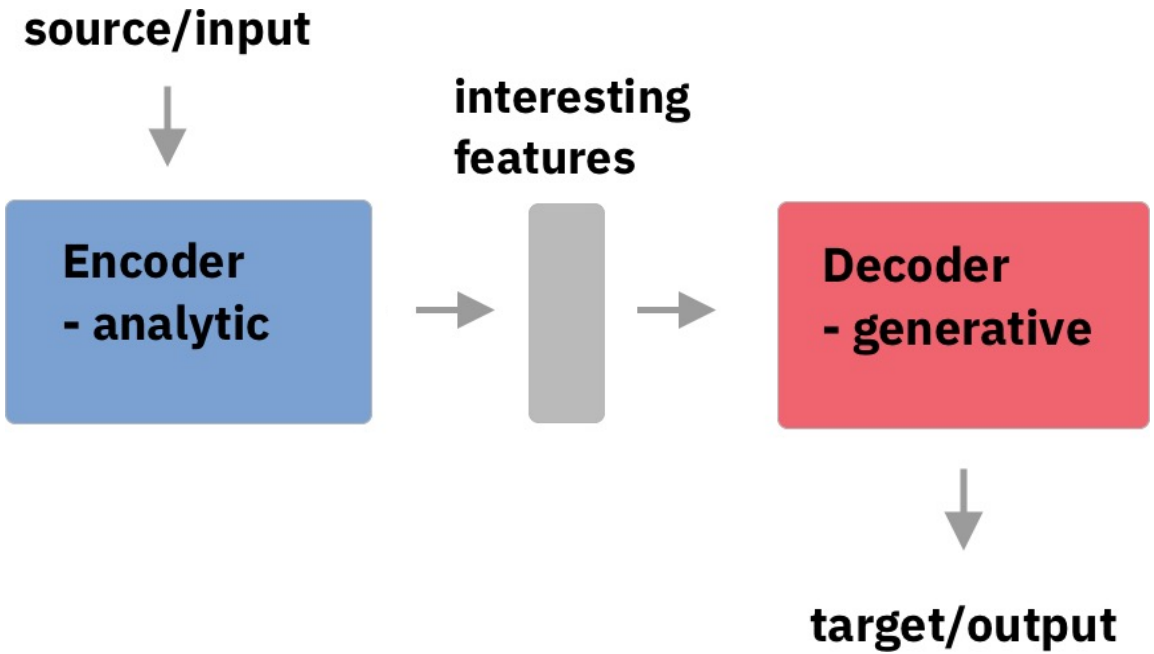
# Atoms as letters, molecules as words



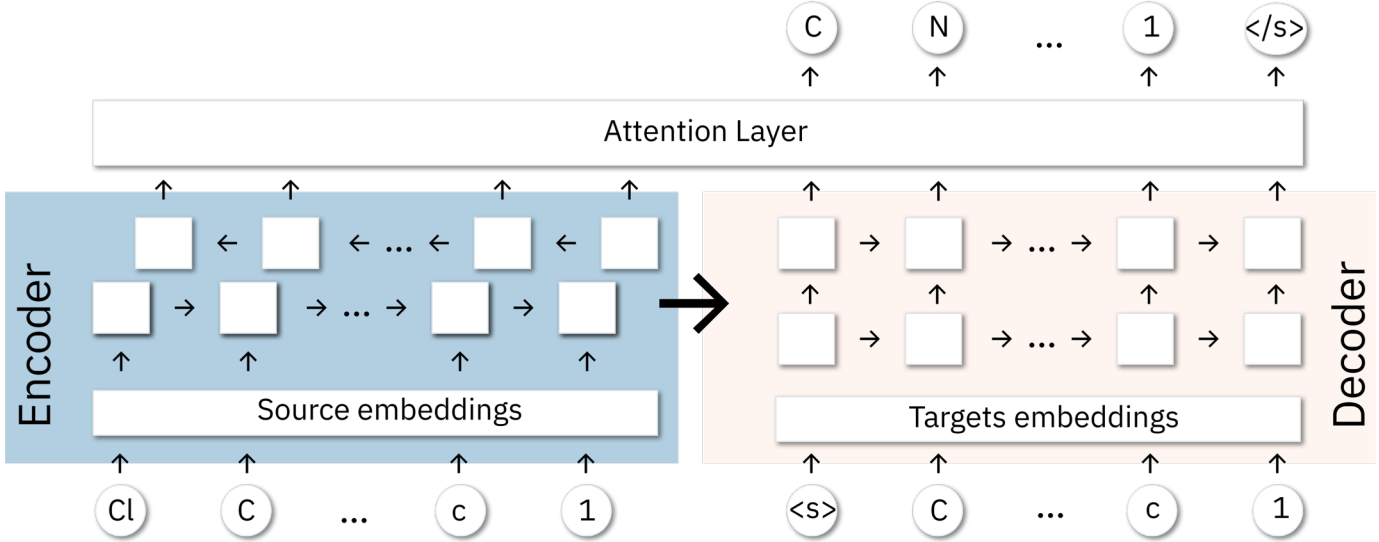

**SMILES to SMILES** prediction
with sequence-2-sequence models

Schwaller et. al. Chem. Sci., 2018, 9, 6091-6098 / Nam & Kim **arXiv:1612.09529**

# How do Seq-2-Seq models work?



# Reaction prediction as Translation problem



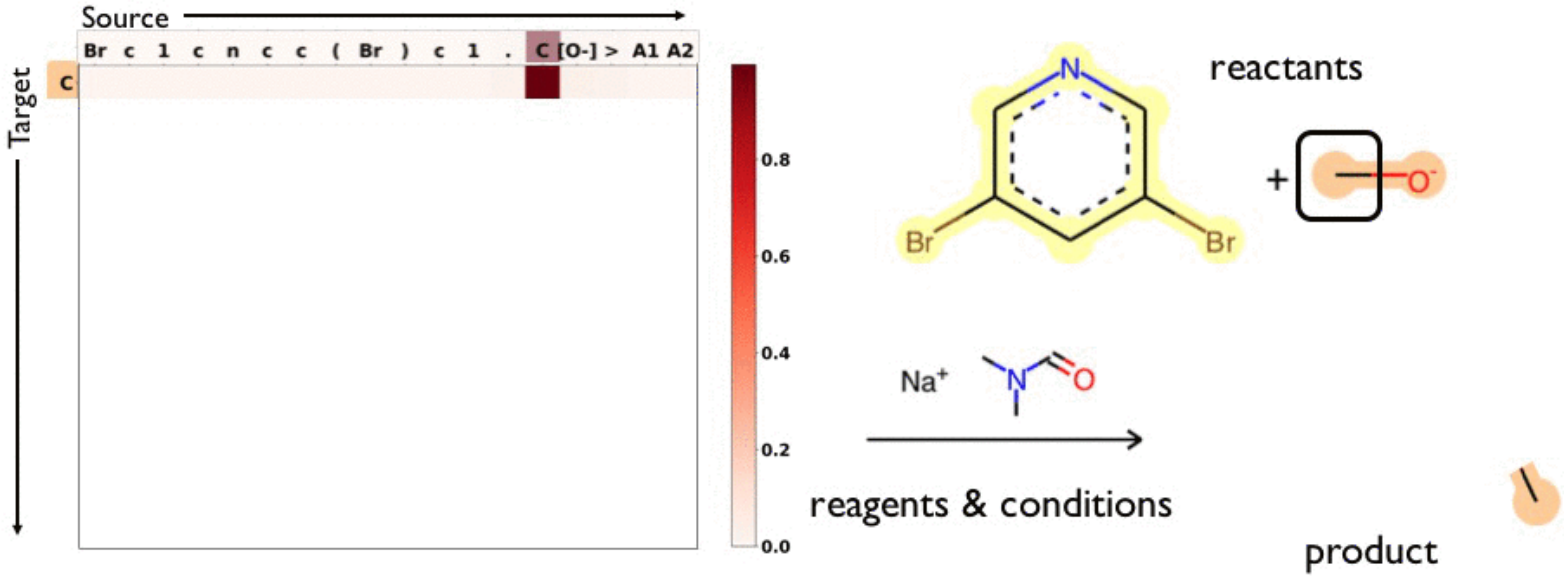

- Fully data-driven
- Template free,
- Trained end-2-end
- NO prior chemical knowledge incorporated into the model

# Open Datasets

| | | | |
|---|---|---|---|
| Lowe2017 | 1,088,170 | reactions | noisy, hard, stereochemistry |
| NIPS2017 (subset) | 479,035 | reactions | cleaned, simplified |

**Lowe, D.** Chemical reactions from US patents (1976-Sep2016) (2017). URL https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.
**Jin, W., Coley, C. W., Barzilay, R. & Jaakkola, T.** Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. In *NIPS* (2017).

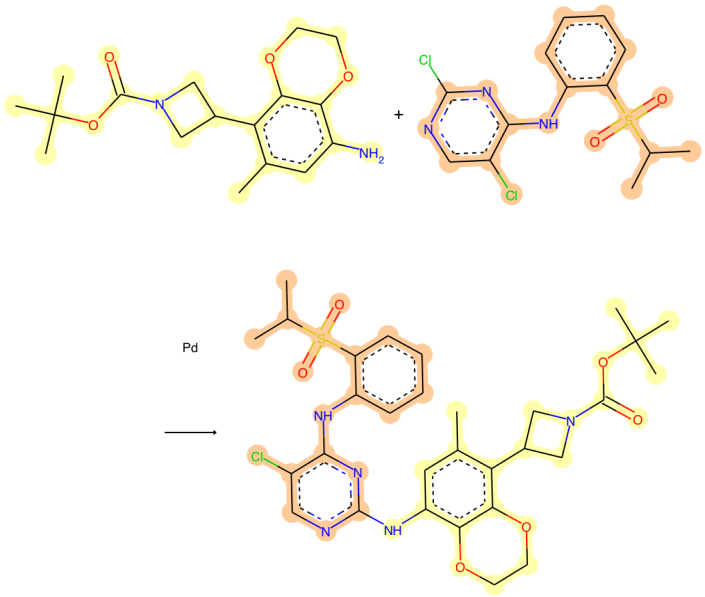
# Attention weights




Plotting the attention weights at every decoder time step.

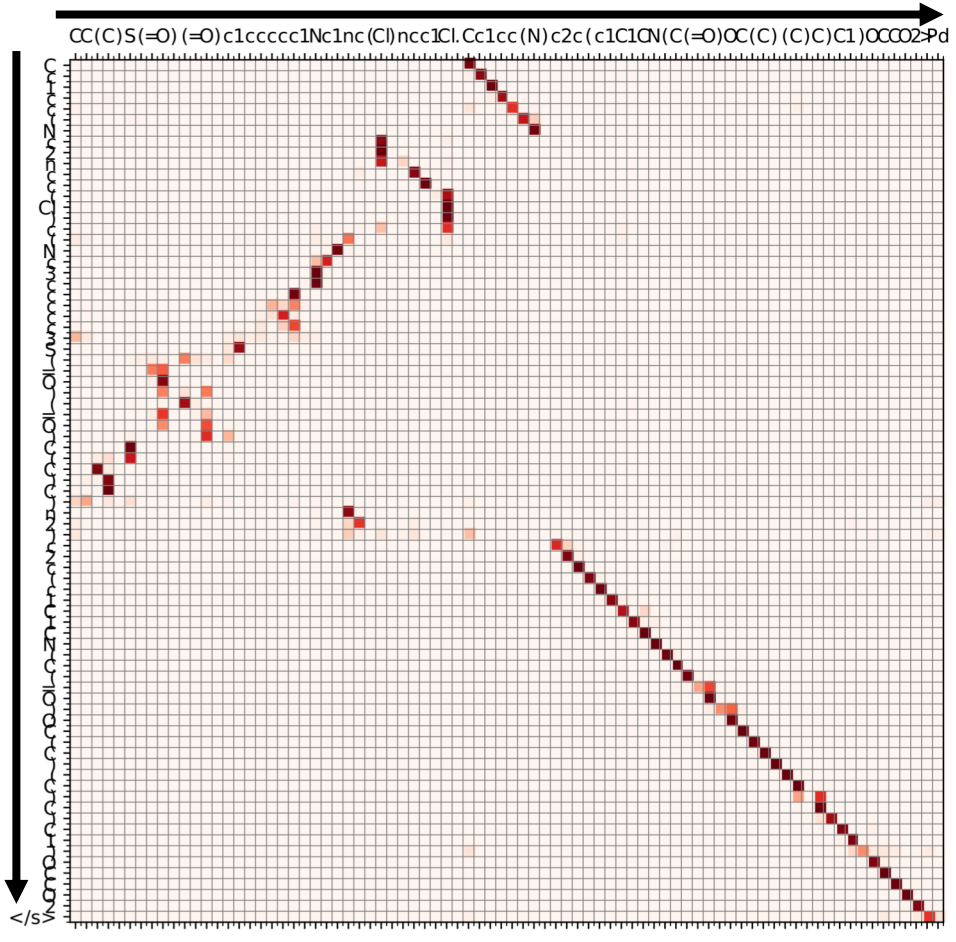
# Attention weights

Chloro Buchwald-Hartwig amination:
**US2017**0240534A1




Source: Reactants

Target: Product



Fully data-driven

Trained end-2-end

Template-free

**SMILES-2-SMILES**

Attention weights and confidence score

No chemical knowledge incorporated

Less than 0.5% invalid SMILES

No atom-mapping required

Only as good as the training data

Difficult to include negative data

**Limitations**

Chemically unreasonable predictions possible

# What other **template-free** approaches exist?

## Graph neural networks

- Jin et al. (MIT, 2017): Weisfeiler-Lehman Networks (WLDN)
  - Network 1: **Reaction center identification**
  - Network 2: **Product candidate ranking**
  - Trained separately
  - Outperforms template-based by 10% on USPTO_500k dataset
- Bradshaw et al. (Cambridge, 2018): Electron path prediction
  - Gated Graph Neural Networks
  - Outperforms WLDN on USPTO_350k dataset
    (subset without more difficult reactions, e. g. cycloadditions)

**Fundamental limitation: require atom-mapped training sets**

# How do we perform compared to others?

Reactants > reagents → Products

Top-1 accuracy:

| Data set | Jin et. al., WLDN | Schwaller et. al., Seq-2seq | Bradshaw et al., GGNN | Our new model |
|---|---|---|---|---|
| MIT_500k | 79.6 % | 80.3 % | | **90.4 %** |
| -subset **350k** | 84.0 % | | 87.0 % | |

Schwaller et al.: Chem. Sci., 2018, 9, 6091-6098
Jin et al.: NIPS, 2017, 30, 2607-2616
Bradshaw et al.: arXiv:1805.10970

# How do we perform compared to others?

Reactants & reagents mixed → Products

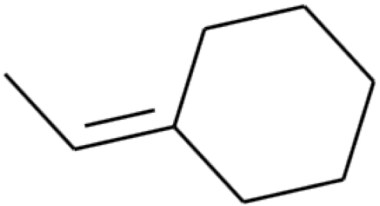
| Data set | Jin et. al., | Schwaller et. al. | Our new model |
|---|---|---|---|
| MIT_500k | 74.0 % | <74.0 % | **88.6 %** |

On USPTO_MIT benchmark dataset.

Schwaller et al.: Chem. Sci., 2018, 9, 6091-6098
Jin et al.: NIPS, 2017, 30, 2607-2616
Bradshaw et al.: arXiv:1805.10970

# What else can we do with AI models ?

# Reaction scoring

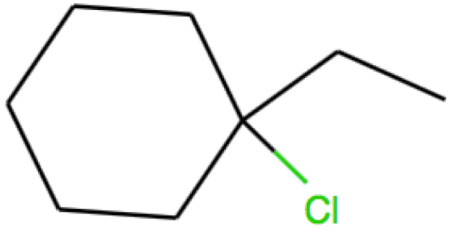


+ HCl

Markovnikov

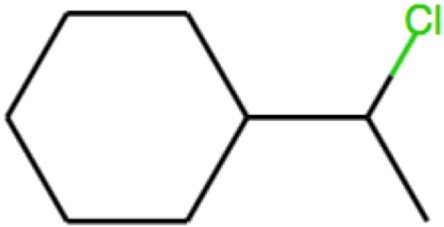
Anti-Markovnikov

CC=C1CCCCC1.Cl>>CCC1(Cl)CCCCC1




Score: 0.99

CC=C1CCCCC1.Cl>>CC(Cl)C1CCCCC1




Score: 0.001

When we provide the model with **reactants>>products**

# What about retrosynthesis?

products → reactants

<RX_1> C O C ( = O ) C C C ( = O ) ....

C 1 = C O C C C 1 . ... c 1 c c c ( O ) c c 1 O

| USPTO_50k | Top-1 [%] | Top-2 [%] | Top-3 [%] | Top-5 [%] | Top-10 [%] | Top-20 [%] | Top-50 [%] |
|---|---|---|---|---|---|---|---|
| baseline | 35.4 | | 52.3 | 59.1 | 65.1 | 68.6 | 69.5 |
| Liu et. al. | 37.4 | | 52.4 | 57.0 | 61.7 | 65.9 | 70.7 |
| **Our model** | **54.0** | **65.6** | **69.9** | **74.0** | | | |

Liu et al. *ACS Cent. Sci.* 3, 10, 1103-1113

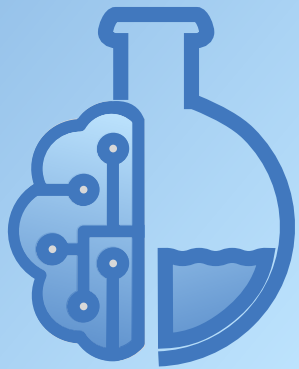
Retrosynthesis model products → Top-1 Reactants (not unique) → Prediction model → **75%** predicted products equal starting products

Trained on Pistachio

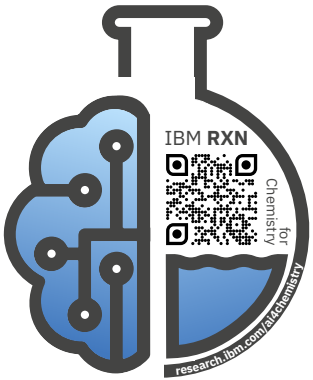
# IBM **RXN** for Chemistry

#RXNFORCHEMISTRY

Freely available now:
**research.ibm.com/ai4chemistry**



# AI model Publications




Molecular Transformer for Chemical Reaction Prediction and Uncertainty Estimation

PUBLICATIONS

by: Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Costas Bekas, Alpha A Lee

Read more ›

November 14, 2018

"Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models

PUBLICATIONS

by: Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, Teodoro Laino

Read more ›

August 9, 2018

Pre-prints available on rxn.res.ibm.com

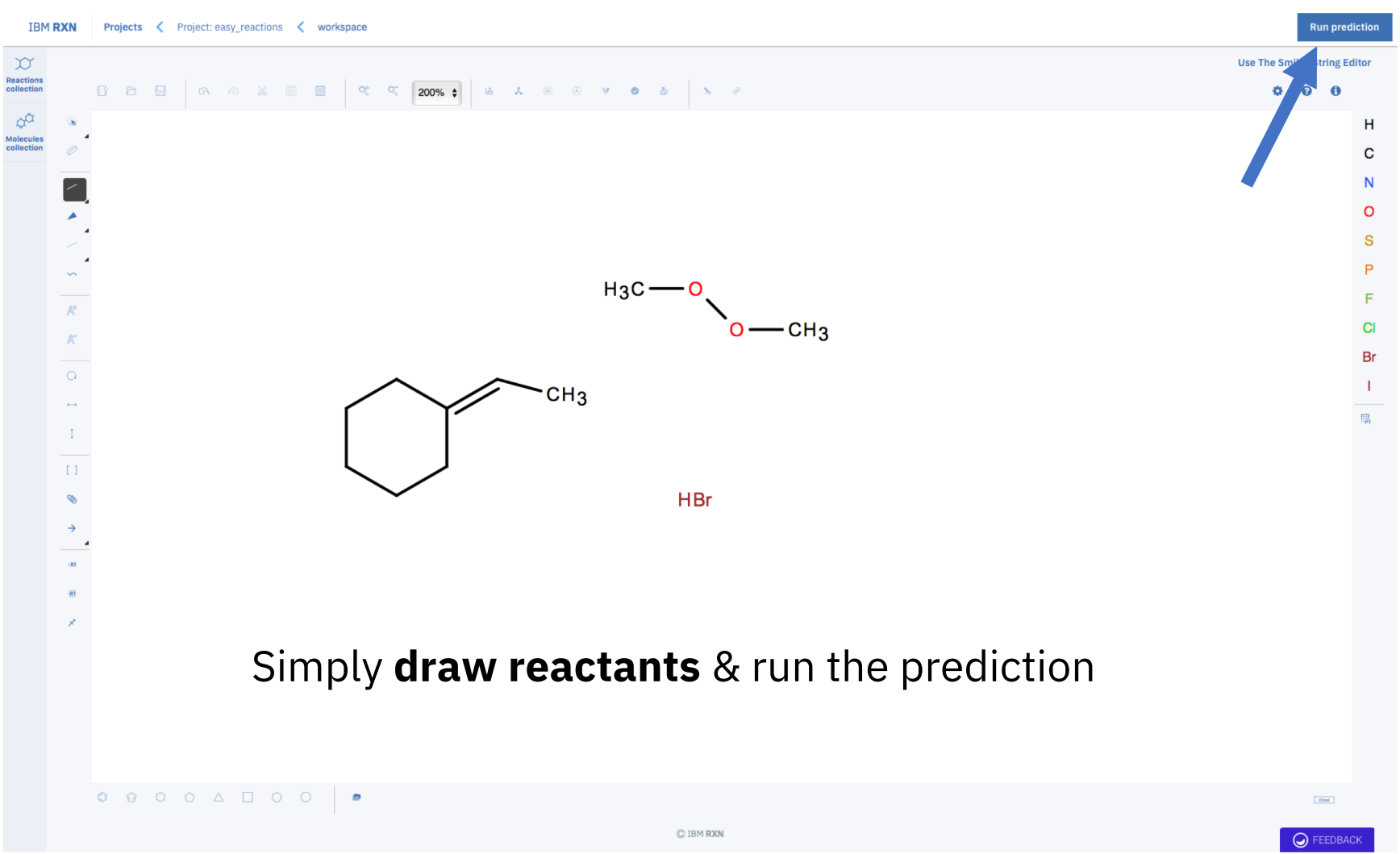
teo@zurich.ibm.com

@teodorolaino

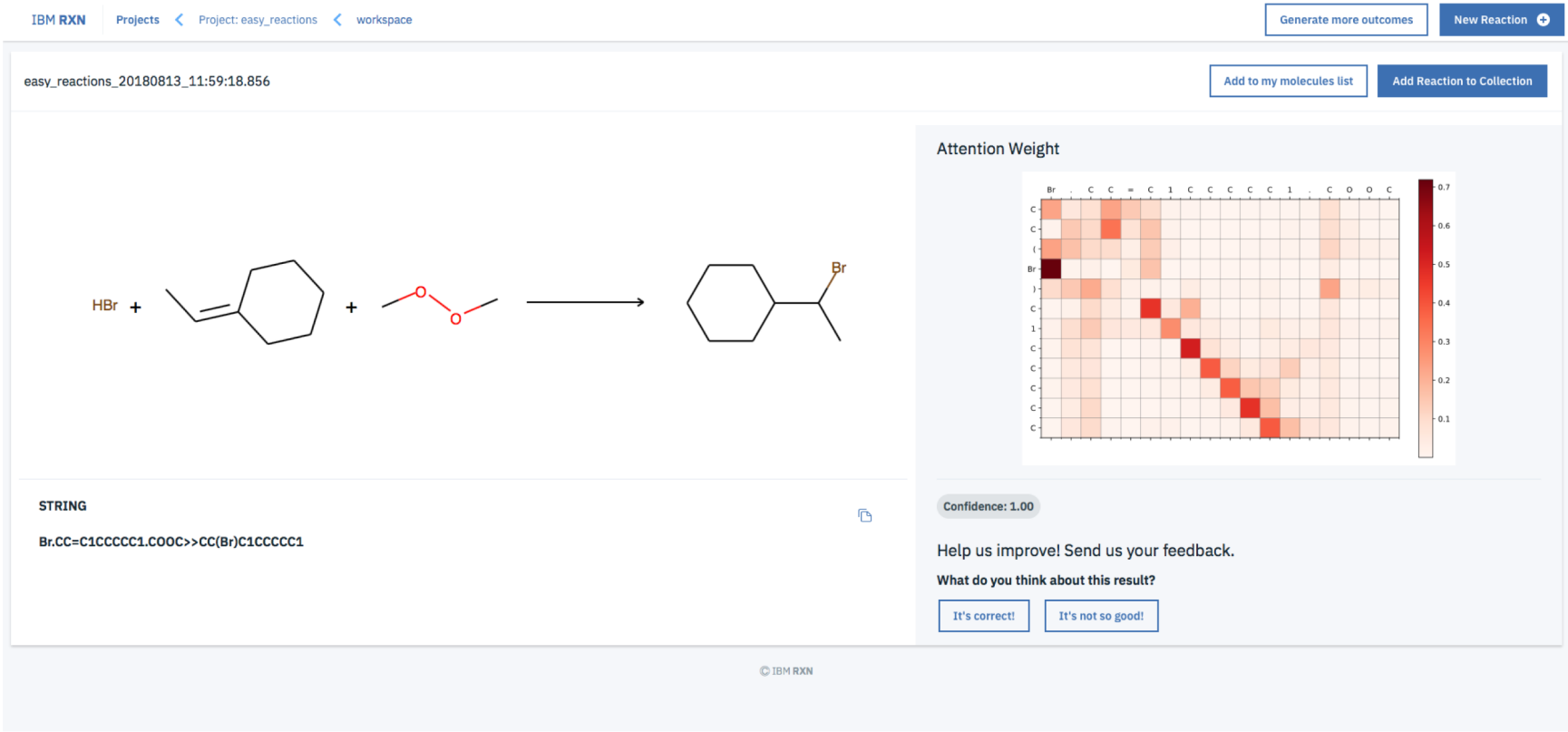
# Questions?


Teodoro Laino

teo@zurich.ibm.com / @teodorolaino




Simply **draw reactants** & run the prediction




Get back the **product**, the **attention** weights and the **confidence**